Minimum Viable Experiment to Replicate

Berna Devezer[1,2] **and Erkan O. Buzbas**[2]

[1]Department of Business, University of Idaho
[2]Department of Mathematics and Statistical Science, University of Idaho

In theory, replication experiments purport to independently validate claims from previous research or provide some diagnostic evidence about their truth value. In practice, this value of replication experiments is often taken for granted. Our research shows that in replication experiments, practice often does not live up to theory. Most replication experiments involve confounding factors and their results are not uniquely determined by the treatment of interest, hence are uninterpretable. These results can be driven by the true data generating mechanism, limitations of the original experimental design, discrepancies between the original and the replication experiment, distinct limitations of the replication experiment, or combinations of any of these factors. Here we introduce the notion of *minimum viable experiment to replicate* which defines experimental conditions that always yield interpretable replication results and is replication-ready. We believe that most reported experiments are not replication-ready and before striving to replicate a given result, we need theoretical precision in or systematic exploration of the experimental space to discover empirical regularities.

Replication | reproducibility | epistemic iteration | exploratory experimentation | idealized experiment | minimalism

Correspondence: ***bdevezer@uidaho.edu***

Introduction

In "What is good mathematics?," mathematician Terence Tao presents a number of scenarios that represent how fields can stagnate (1). Two of these scenarios appear to be of particular relevance to some subfields in social and behavioral sciences that have struggled with a putative replication or reproducibility crisis in the last decade and a half:

- "A field which becomes filled with many astounding conjectures but with no hope of rigorous progress on any of them",
- "A field which now consists primarily of using ad hoc methods to solve a collection of problems which have no unifying theme, connections, or purpose".

Tao's subsequent observation that in mature and well-developed fields, the earlier reliance on heuristics and lack of rigor should be replaced with systematic, programmatic, rigorous theoretical investigation to avoid stagnation appears to generalize outside of mathematics. This observation is parallel to Lakatos's *methodology of scientific programs* where scientific research programs evolve over time by responding to earlier problems and limitations, and theories are modified to accommodate anomalies and address problems (2, Ch. 6.2). Progressive programs gain more explanatory and predictive power over time. However, some fields in social and behavioral sciences have not evolved as such in rigor as they have matured in age (e.g., social psychology). Astounding conjectures and *ad hoc* methods have become solidified as norms in predominant scientific paradigms of these fields, leading to the state of affairs that is commonly referred to as the replication or reproducibility crisis. In Lakatos's framework, these may exemplify degenerate research programs that have become stagnant and are unable to deal with anomalies. A scientific reform movement[1] has emerged against this backdrop that has centered around the ideas of replication of experiments and reproducibility of scientific results. The question arises: Can replication experiments help isolate the truth value of results and promote epistemic progress in fields that are potentially teeming with unverifiable findings (4, 5) and are often characterized by lack of clear theoretical progress (6, 7)?

In this paper, contrary to the expectations internalized by the reform movement, we argue that replication experiments are not well situated to improve the theoretical or empirical rigor in fields that have become stagnant for reasons described by Tao[2], specifically when employed in a diagnostic capacity. To the contrary, we believe that most experiments are not replication-ready and that replication experiments might need to be based on rigorous exploration of experimental space or theoretical advances to yield meaningful diagnostic results. We seek to delineate the characteristics of experiments required for replication-readiness and introduce the notion of the *minimum viable experiment to replicate*, contrasting it with the notion of experimentum crucis in theory-driven fields.

Before we advance our core argument, it is necessary to define a replication experiment and identify its aims. Only then will we be able to elucidate what cannot be achieved by replicating experiments that are not replication-ready and why. We start with an overview of our research program wherein we have worked toward developing a mathematically grounded account of replication and reproducibility of scientific results.

[1]While this movement has generated heterogeneous communities and efforts, it has often been referred to as a single, unique entity instigated by the replication crisis (3). Out of expediency, we use the phrase "scientific reform movement" to refer to the totality of the reforms and activities surrounding replications and reproducibility, while acknowledging that this is a heterogeneous body of literature.

[2]While our argument regarding replication experiments holds in general, the reason we are writing this paper is particularly because their role is exaggerated and their epistemic status is elevated in fields currently looking for solutions to a host of empirical and theoretical problems. We will attempt to expound that it is misguided to expect replication experiments to achieve many of the outcomes expected from them in fields undergoing a "crisis".

Replication and reproducibility

While many different definitions and interpretations of the term *replication* exist (see 5, for an overview), a common way it is understood is the redoing of a whole experiment by repeating its methods, procedures, and analyses to observe whether the replication will successfully produce the same or sufficiently similar results as the original (e.g., 8–10). This view is consistent with the Mertonian account which delineates the central premise of replication as separating true from false claims (11). Many replication advocates in science reform appeal to this *diagnostic* value and the *demarcation* function of replication (12). The diagnostic ideal about replications can be traced back to Popper who suggested that reproducibility of experimental results should be a basic methodological requirement for establishing the validity of scientific claims (13).

In our research program, we have attempted to lend mathematical formalism and precision to the concept of replication (14, 15), particularly because a mathematical formalism is necessary for statistical evaluation of results that ultimately come from data. Such formalism also allows us to study limitations of the concept of replication (16). Because of its centrality in this paper, we reiterate some fundamental components of this formalism here.

Before we can define a replication experiment, we need to introduce the notion of an *idealized experiment*—a theoretical entity that captures the structure of a scientific experiment as it bears on statistical evaluation of results. Assuming some background knowledge K on a scientific phenomenon, a scientific theory makes a prediction in principle testable using observables, the data D. A scientist formulates a mechanism generating D under uncertainty and represents it as a probability model M including its assumptions. Given D, the scientist is interested in performing inference on some unknown part of M. To assess to which extent the desired inference is confirmed by D, the scientist uses a fixed and known collection of methods S evaluated at D. This description captures some key components of studies whose population characteristics can in principle be tested. We break down S into two components: S_{pre} and S_{post}. S_{pre} comprises the set of *scientific* methodological assumptions preceding data collection and procedures implemented to obtain D. More specifically, S_{pre} captures the premises underlying the design and execution of an experiment such as procedures, instruments, and manipulations. S_{post} comprises the set of *statistical* methods applied on D to obtain the result of interest.

The idealized experiment is the (ordered) tuple

$$\xi := (K, S_{pre}, M, S_{post}, D).$$

Two components need further clarification. D consists of random values and the data structure. This implies that ξ is random (if a particular fixed realization of D is needed, it is made explicit). K, which carries the state of scientific knowledge on the phenomenon of interest—including theoretical assumptions, cultural and historical context, experimental paradigms, and the scientific language—used to conceptualize, design, and perform the experiment is kept implicit.

Invoking the concept of a *result* from ξ is not necessary to discuss properties of ξ. However, the goal of conducting ξ in practice is to perform inference about some aspect of the real world. Statistical inference comes as a *decision* in the form of a result from ξ. A result, $R(d, c)$, then can be satisfactorily defined as a function of d, a user-defined decision rule (with respect to a user-defined loss function), and possibly of c, some fixed user-defined criterion.

We define a *replication experiment*, ξ', as a specific type of idealized experiment that aims to reproduce a given result from ξ by generating independent D. ξ' is an *exact* replication[3], if

$$\xi' := (K', S_{pre}, M, S_{post}, D').$$

That is, an exact ξ' differs from ξ only in K' and D'. Crucially, $K' \supset K$, because components K, S_{pre}, M, S_{post} must be passed onto ξ' as part of K', so that ξ' is informed of what exactly it is expected to replicate. D' are always generated independently and randomly in an attempt to reproduce the original result. If ξ' differs from ξ except in K and D in a way to disrupt the inferential equivalence of the components, it is a *non-exact* replication. If ξ' confirms a targeted result of ξ, it is said to have reproduced that result[4]. The reproducibility rate then is the relative frequency of reproduced results in a sequence of ξ'. One of our major theoretical results is that any given sequence of exact or non-exact replication experiments converges on a true reproducibility rate, as a function of the components of the original experiment (16). That is, the true reproducibility rate of a result is not only a function of the true data generating mechanism, but also a function of K, M, S, D. Further, depending on the degree of non-exactness, the estimated reproducibility rate from a given sequence may or may not be close to the true reproducibility rate of the original result. These definitions and theoretical results come in handy to advance our current argument.

Limitations of replication experiments

Limitations of replication experiments have been discussed in the literature. Three issues stand out:

1. Designing and performing exact replication experiments is notoriously difficult and oftentimes, practically impossible.

2. Even when possible, inferences that can be made from exact replication experiments are narrow and do not fulfill the prescribed diagnostic function.

3. As common as they are, non-exact replication experiments are limited in what they can achieve due to the

[3] Sometimes referred to as *direct* replication.

[4] For clarity, we use the term "reproducibility" exclusively to refer to whether an experimental result can be reproduced in a replication experiment—that is, results reproducibility. Sometimes the term replicability is used in the literature to this effect. For internal consistency within our research program, we prefer to continue using the former terminology.

openness of conceptual scope and the range of potential confounds introduced. Particularly, inference afforded by non-exact replication experiments is not a good proxy for inference desired from exact replication experiments.

The first point has been considered largely trivial and widely acknowledged by both proponents (e.g., 17, 18) and critics (e.g., 9, 12, 19) of replication. Our theoretical framework on idealized experiments makes clear why. For a replication experiment to be considered exact for the purposes of reproducing a single result of an original experiment, the replication: (i) has to assume an equivalent model to that assumed in the original, (ii) has to repeat equivalent experimental and analytical procedures, and (iii) has to have an equivalent data structure (including, for example, the sample size[5] and sampling from the same population). In practice, such standards might prove impossible to attain, but they are relevant. These standards establish a mathematically well-defined reference point against which we can compare the practical outcomes as gold standard. For example, the original experiment may not explicitly report the background knowledge which undergirds the study (see 16, p. 12, for an example) or may not completely report all methodological procedures or non-trivial decisions. Or some of the background conditions impacting the original experiment may not exist at the time of the replication experiment. The replication experiment may also be subject to different resource constraints (e.g., access to a different population) and may have to make different design choices (e.g., changing technological standards in experimental procedures). As a result, the best that can be achieved in practice is some level of similarity between the original experiment and its replication (17, 19). To assess the value of a replication mathematically in a consistent manner, we should always fall back to the case of an exact replication that has the same reproducibility rate as the original experiment.

The second point is nontrivial. Bogen (20) refers to the diagnostic ideal about replications as the *received doctrine about replicability*[6] and criticizes the normative claim that existing scientific claims can only be legitimately confirmed or disconfirmed by providing empirical evidence that they are/can be reproduced. Bogen (20) argues not only that corroborating evidence in support of a claim can come in different forms but also that replication experiments may not even be able to provide relevant resolution. He meticulously investigates case studies from neuroscience and medicine, documenting epistemic progress via irreproducible results and concludes:

> "Repeated applications of the same experimental or observation procedure typically do not, and are not expected to, produce exactly the same results. Although replications of the relevant procedure are required, a result obviously does not need to be thrown out because it differs from previous results." *(20, p. 22)*

The received doctrine about the role of replications is too narrow[7] and unrealistic when juxtaposed against how scientific progress is made in practice.

In a simulation study, we created a stochastic model of the scientific process and observed how a community of agents pursuing different scientific strategies searched for a true model (21). In many scenarios, scientists were able to make true discoveries but unable to reproduce them in replication experiments and in others, they were able to reach 100% reproducibility without ever converging on the true data generating model. There did not appear to be a meaningful correlation between scientific discovery and reproducibility of a true result, providing evidence against the diagnostic ideal, in line with Bogen's observations.

Why then do exact replication experiments fail to discriminate between true and false results? Our theoretical framework of idealized experiment provides a clue and the answer is multifaceted:

- In any scenario including making inference from a sample to a population, there is uncertainty we cannot eliminate due to sampling variability. This is true for any given experiment, original and replication alike. Even an exact replication experiment is not equipped to eliminate uncertainty due to sampling variability. In fact, the idea of *eliminating uncertainty* is a misinterpretation of the goals of statistical inference. Statistics' goal is to *quantify uncertainty*, by acknowledging its presence, as opposed to eliminating it.

- The reproducibility rate of any given experimental result varies not only as a function of the true data generating mechanism but also the components of the idealized experiment (16, 22). Experiments are subject to many other sources of unaccounted for uncertainty than sampling variability due to decisions regarding specification of scientific and statistical models, sampling scheme, methodological procedures, and other design elements. Feest (9) refers to these problems as systematic error, and argues that these conceptual and material presuppositions and uncertainties cannot be remedied by exact replications. She observes that many exact replications are focused on ruling out sampling error[8] (which is already not possible by the

[5] Sample size is not trivial. An experiment planning for sample size n might get sample of size m, such that $m < n$ due to, for example, missing data. Missing data changes the information available from true mechanism generating the data, and hence the distribution of a statistic that is used to draw statistical conclusions.

[6] The term replicability is used in the same manner as we use results reproducibility or reproducibility here.

[7] Bogen keeps his discussion of the received view focused on exact replications and does not explicitly examine non-exact replications.

[8] One of the editors of the volume questioned whether this characterization might be a straw man. We believe not. This alleged ability to control for or rule out sampling error has been acknowledged explicitly in discussions of the functions of replications (for a well-cited example, see 23). Moreover, this is apparent in the post-replication crisis language used to evaluate many non-replicable effects in the literature. For example, see a blog post by social psychologist Michael Inzlicht (24) wherein he reviews replication evidence for the ego-depletion phenomenon and concludes "the work upon which our celebrated theory was based was not replicable, not real". Similarly, referring to the state of social priming literature in the aftermath of the replication crisis, Chivers (25) says: "it became clear that many of the prob-

first point) and ignore much of this systematic error. When an original study carries such issues of conceptual scope and violation of assumptions, its replication remains undiagnostic as to whether its results are to be attributed to the phenomenon of interest or the features of the experiment. For example, it is a statistical fact that model misspecification (e.g., measurement error) might render true results less reproducible and false results more reproducible (15, 21). In fact, statisticians have paid special attention to various ways to misspecify a model, especially in the last 20 years. As such, "even if direct replication can confirm the existence of an effect, it cannot say what kind of effect." (9, p. 899).

- Reproducibility rate of a given (true or false) result also varies with the decision rules we use to identify what counts as a result and how we determine whether it is reproduced (15). We may say that a result is successfully reproduced if the effect observed in the replication experiment is in the same direction as in the original experiment or only if the effect size estimate from the replication experiment falls within some small standard errors around the original point estimate. The first rule imposes a less severe constraint and would be expected to result in a higher reproducibility rate than the latter.

In short, even exact replication experiments are subject to multiple (scientific, operational, and statistical) sources of uncertainty, limiting their usefulness and diagnosticity regarding the truthfulness of scientific claims under study [9].

In light of the above discussion, the third point becomes less surprising. Where even exact replication experiments fail to provide clear empirical evidence in support of or against a scientific claim, what can be accomplished by a haphazard sequence of non-exact replication experiments will prove impossible to pin down[10]. A common argument goes that how close or similar the methods and procedures used in a replication experiment are to those used in the original experiment is representative of the quality or diagnostic value of replication (8, 11, 18). Similarity or closeness in this regard, however, is difficult to define since it has to be with respect to a reference standard (9) and even more difficult to measure. In Buzbas et al. (16), we made progress toward providing a formalized definition for some components of the idealized experiment but the latter aim remains elusive. For example, K of the replication experiment has to differ from that of the original experiment, because it has to carry over the result obtained from the original experiment to assess whether it is reproduced. However, if the tacit, implicit aspects of the cultural, social, scientific, paradigmatic assumptions underlying the experiment cannot be completely transferred due to a lack of transparency or even a lack of awareness of original scientists, the replication is likely to be non-exact and nontrivially different from the original. This component of the idealized experiment is still difficult to define with precision and to repeat or emulate. Conditional on an inferential goal, we can say more on what similarity means regarding the remaining components of the idealized experiment. For example, an exact or close replication experiment conditional on the inferential goal is possible to reproduce a result, if there exists an isomorphic transformation between the models assumed in the original and the replication (result 4.2, 16). That is, they need to be equivalent with respect to the likelihood function (i.e., likelihood principle) to draw identical inferences about unknown quantities. Similarly, pre-data methods, statistical methods, and data structure do not need to be identical in every sense to reproduce a result (results 4.3-4.6, 16). Equivalence of components for purposes of reproducing a result is easier to achieve for statistical methods and data structure, and more difficult with pre-data methods (e.g., experimental instruments, procedures, operationalization of variables). In a series of interviews with reviewing editors at the journal Science, Peterson and Panofsky (11) observe that especially in fields that have high task uncertainty, that is, where experimental practices, procedures, and techniques are either unstandardized or unstandardizable, replication experiments tend to be piecemeal and undiagnostic. This situation is reminiscent of Tao (1)'s second scenario regarding less settled fields relying on ad hoc methods to solve idiosyncratic problems. As pre-data methods begin to diverge between experiments, replications quickly run into issues of conceptual scope, as highlighted by Feest (9). In regular scientific practice in many such fields with high task uncertainty, the formal equivalence between these unstandardized components is rarely established and most decisions of similarity or closeness is grounded in scientists' intuition, which results in non-exact replications becoming the norm where the degree of non-exactness is unknown and unmeasurable.

In Buzbas and Devezer (28), we present some theoretical progress toward providing a formalized definition and a formal measure of the distance between non-exact experiments. Our distance measure relies on the mathematical observation that reproducibility rates vary with changes in experimental components. While this theoretical advancement is promising in informing our choice of non-exact experiments to perform in a sequence to satisfy specific epistemic aims (e.g., eliminating assumptions or finding out whether empirical re-

lematic findings were probably statistical noise – fluke results garnered from studies on too-small groups of people". In one of the most influential papers of scientific reform, Wagenmakers et al. (26) explicitly states: "Research findings that do not replicate are worse than fairy tales; with fairy tales the reader is at least aware that the work is fictional." Such statements assume that replication experiments should be capable of ruling out statistical flukes, and thereby sampling error.

[9] We acknowledge that there are different proposals regarding the primary function of replication experiments. For example, Fletcher (27) suggests that data obtained from replication experiments is best suited for meta-analytic evaluation rather than to assess the validity of an original finding. Our target is specifically this latter, diagnostic role attributed to replications in the current paper. We advance an argument to show what it takes for a replication experiment to satisfy that diagnostic function and our argument is agnostic regarding these other functions or uses of replications.

[10] In Buzbas and Devezer (28), we chart a systematic approach to a planned sequence of non-exact replications to help map the parameter space. Similarly, Bogen (20), Burian (29), and Steinle (30) show historical examples of non-exact experiments being used for exploration and evidence triangulation. Our argument regarding non-exactness here regards their narrow use as a proxy for exact replications, in a confirmatory fashion.

sults are robust to changes in specific experimental components), it does not help bridge the gap between experimental distance and diagnostic value of evidence obtained from non-exact experiments. By definition, the larger the distance between experiments, the more the reproducibility rates will vary, potentially leading to non-replicable results regardless of any underlying true regularities.

The problem with non-exact replications is one of underdetermination (31)[11]. As we reasoned earlier, even exact replication experiments are underdetermined in which their results cannot be singularly attributed to the phenomenon being investigated. Non-exact replication experiments introduce even further confounds and potential causes to which results can be attributed to, exacerbating the problem of underdetermination. There's no way of knowing whether a "successful" or "failed" replication simply repeated a systematic error preexisting in the original experiment, ran into an instance of sampling error, introduced a new systematic error via different experimental components whose equivalence has not been established, or actually reported something regarding the truth of a scientific phenomenon. The reproducibility rate estimated based on a sequence of non-exact replication experiments converges to the mean reproducibility rate of results from all experiments,[12] as opposed to the true reproducibility rate of the original result of interest (result 5.1, 16). Even if the original experiment has captured a false result that has a reproducibility rate close to 0, we can easily run a sequence of seemingly close non-exact replications that yield results that are reproducible 80% of the time. Ultimately, the observed pattern of results in non-exact replications is multiply determined and the specific causes remain unidentifiable based on replication results alone.

To summarize, replication experiments are not generally fit to accomplish common aims often attributed to them such as isolating signal from noise (32), excluding or exposing unlikely results (13), or simply separating true results from false ones—the diagnostic ideal. What, then, are replication experiments capable of? Potential answers include but likely are not limited to:

1. Gradually increasing evidence in support of an original result and the precision in our inference[13],
2. Estimating the reproducibility rate of a given experimental result, if that is of particular interest—as used in quality control applications,
3. Systematic exploration of the experimental space via planned non-exact replications,
4. Performing theory-guided robustness tests via non-exact replications[14].

The extent to which these outcomes can be achieved will greatly depend the properties of the original experiment and its replication. Setting these functions aside, we turn to our core argument concerning the diagnostic readiness of replication experiments.

Replication-readiness of experiments

In many experimental fields individuation judgments involved in the choice of experimental components result in high levels of epistemic and methodological uncertainty (9, 22)[15]. These choices lead to many auxiliary hypotheses that cannot be decoupled from the scientific hypothesis (31). For example, Deaton and Cartwright (34)'s extensive investigation of randomized controlled trials (RCTs) reveals the limitations of this well-regarded form of experimentation. They advance a convincing argument against the notion that the average treatment effect (ATE or result, for our purposes) estimated "from an RCT is automatically reliable, that randomization automatically controls for unobservables, or worst of all, that the calculated ATE is true" (34, p. 29). Confusing statistical inference with scientific inference is usually the underlying problem and the drivers of these inferential errors can be traced back to the components of the idealized experiment: missing background knowledge, violated model assumptions, measurement error, imprecise treatments, nonrepresentative samples, heterogeneity of treatment effects across sub-populations, invalid statistical methods (also see Geyer,

[11] This is true for all experiments. However, for replications specifically underdetermination is augmented by the distance between the original and the replication experiment, and potential confounds introduced by the nature of this distance. A replication experiment in its diagnostic capacity is not simply tasked with testing a specific hypothesis but with confirming the evidence presented in an original experiment with respect to a specific hypothesis. Hence the source of underdetermination is not just the detachment of empirical evidence from theoretical prediction but also the divergence between original and replication experiments.

[12] This is a consequence of the Law of Large Numbers. For a fixed result, the mean reproducibility rate of the result obtained from a randomly chosen subsequence of experiments with different reproducibility rates is equal to the mean of reproducibility rates of these experiments.

[13] This is different from the diagnostic ideal in two ways: First, it focuses on precision and not accuracy. Second, it is about continuous accumulation of evidence—similar to Fletcher (27)'s meta-analytic perspective of replications—rather than a one-time decision about validity of experimental results.

[14] We believe non-exactness can be used to test for robustness only if the set of non-exact replications can be strongly linked by scientific theory. Otherwise, evidence from non-exact replications can speak to different theoretical mechanisms. In other words, just because two seemingly similar experiments produce results that appear to be compatible with each other, the inference that they are driven by the same mechanism is not warranted and the aggregate evidence may not speak to the robustness of an underlying effect. There needs to be an external way to connect the two empirical demonstrations theoretically to speak of robustness (see 33, for more on epistemic iteration and generalizability) .

[15] Feest (9) calls the sum of such individuation judgments *conceptual scope*. Conceptual scope of an experiment basically refers to all assumptions already built into the experimental design by choice of independent and dependent variables as well as all judgments regarding the identification of the relevant features of those variables. In Devezer et al. (15) and Devezer and Buzbas (33), we use an expanded interpretation of such scope in the sense of a set of experimental models that can be accommodated within the boundaries of a theoretical subsystem under study as defined by a theoretical model. The less well defined the boundaries, the more open the conceptual scope and the larger the number of potential experimental models that may loosely relate to the theory. There is usually a one-to-many mapping between a well-defined theoretical model and experimental models associated with it. Open conceptual scope often implies a many-to-many mapping instead, where any given experiment may be associated with different theoretical models just as closely.

this volume). Deaton and Cartwright further demonstrate how spurious findings obtained in such flawed experiments can be reproduced in exact replications and discuss why in practice well-conducted RCTs that can indeed provide a valid estimate of an ATE are scarce.

If even the gold standard in experimental design is far from yielding interpretable results, most other designs are expected to suffer from similar issues if not many more. As Amrhein and Greenland (22, p.263) observe: "Even the best single studies will be imperfect. In addition to random variation, their results will usually vary from replication to replication because of varying assumption violations, whether recognized or hidden, and thus the observed effect sizes can easily differ across settings." Since "some degree of variation, and hence non-replication, is the norm across honestly reported studies, even when all assumptions are met" (22, p.264), in situations where the assumptions are not met, successful replications should not be expected with any meaningful frequency. For exact replication experiments to achieve their previously mentioned aims, they need to be based on experiments whose components are free from unaccounted for errors (e.g., no measurement error), internally consistent with each other (e.g., data structure is compatible with model assumptions, background knowledge reflected properly in assumed model)[16]. Further, they need to be explicitly and transparently documented to allow for exact replications. As such, most experiments are not replication-ready in the sense that diagnostic, stable inferences from its replications will not be warranted. If we allow that some level of uncertainty is inevitable, is there a way to design replication-ready experiments that might nonetheless yield meaningful, trustworthy inferences?

The standard view of experiment in the twentieth century was theory-driven where the goal of experiment is to test well-defined predictions made by a theory (35). That is, experiments are used to elicit decisive answers to pointed questions raised by theories (36). According to this narrow conception of an experiment, theories should be sufficiently advanced before experimentation can even begin. Experiments are not only inspired and motivated by theory, they are designed, executed, evaluated, and interpreted in light of some theory as well. Newton's *experimentum crucis* was a prime example of such an experiment (37), which was meant to refute wave theories of light (originally appeared in 38). An idealized version of experimentum crucis is a powerful experiment capable of decisively ruling out all other theories that might explain a result of interest except the experimenter's theory (implying a one-on-one mapping between the theory and the experiment). If such an experiment did practically exist, it would take a great deal of theoretical precision, a lack of underdetermination of the experiment by theory, and a narrow conceptual scope. Even Newton's experimentum crucis has not remained unchallenged in this regard (39). Such practical limitations notwithstanding, theoretical maturity and precision could indeed be one path toward designing replication-ready confirmatory experiments. Nonetheless in many theoretically advanced fields, experiments are far from the precision of the experimentum crucis ideal and progress relies on theoretical advancements, model development, and accumulating and triangulating empirical evidence rather than relying on diagnostic replications.

In many areas of social and behavioral sciences, theories are far from any mathematical precision and conceptual scope of experiments tend to be open. Despite the prevalence of using experiments in a confirmatory and seemingly theory-driven fashion, most experimental results are loosely or at times only trivially connected to the theories they purport to speak to (40, 41)[17]. Assuming a field characterized by conditions similar to those exemplified in Tao's hypothetical scenarios, can we talk about replication-readiness? If so, how can experiments attain that status? To this end, we turn our attention away from the standard view[18] and focus on a Baconian variety of experiment as presented by Hacking (ch. 9, 44).

Iterative exploratory experimentation

The Baconian method allows for experiments to be used to explore the world with limited preconception or theorizing about the state of nature. Steinle (30) and Burian (29) coined the term *exploratory experimentation* independently to refer to this variety of experiment aimed at discovering empirical regularities and characterizing phenomena generating these regularities, as opposed to testing theoretical predictions. The epistemic value of this form of experimentation had long been overshadowed by traditional accounts of experimentation and has only in the last couple of decades begun drawing attention.

Exploratory experimentation is said to take place in stages of scientific development where well-formed theories or con-

[16] This simply follows from the nature of performing inference under uncertainty. See Amrhein and Greenland (22) for further discussion on the statistical rationale. In earlier work, we also discuss examples of experimental bias being carried over to replication results, leading to increased certainty in erroneous inferences (15, 16, 21).

[17] The detachment of (statistical) hypothesis testing from scientific theories is not a new realization. Meehl (42) wrote about the lack of cumulative character of scientific knowledge in psychology decades ago. Meehl observed how in "soft" areas of psychology, new theories keep emerging and initially receive a lot of enthusiasm only to die a slow death due to failures in accumulating empirical support. Gigerenzer (43) partially attributes this lack of evidence accumulation to the mindless use of the null hypothesis significance testing rituals that are upheld by collective concession to certain illusions allowing scientists to mistake statistically significant results for highly important discoveries.

[18] Some readers may wonder, as the editors of the issue have, whether we are giving up on the standard view of experiments too prematurely or maybe even haphazardly. We would like to reassure the readers that we do not take this perspective lightly. In fact, we have written about our model-centric view of scientific progress elsewhere (14, 21, 33) and would like to encourage interested readers to follow up on these articles to get a sense for why we believe hypothesis- and result-centric views of science are incompatible with our view. However, we do mean this in a narrow sense. There is no denying that experiments and hypothesis testing have important roles to play in the course of scientific practice. Nonetheless, we believe that they are often over- and mis-used in a confirmatory capacity, especially in disciplines suffering from the previously discussed problems, where the standard view scarcely applies. This is hardly a unique or novel perspective; a model-centric approach to scientific practice is quite common, especially among statisticians.

ceptual frameworks are either nonexistent or deemed unreliable (30). Rather than exclusively referring to specific procedures or individual experiments, exploratory experimentation is characterized by a systematic process of exploration through an elaborate system of interconnected experiments[19]. Across a series of examples from the history of electromagnetism, electricity, organic chemistry, and biochemical research, Steinle and Burian demonstrate that exploratory experimentation has been used by scientists in the formation, stabilization, and formalization of classificatory and conceptual frameworks (29, 30, 35). This exploratory process is characterized by:

- Systematic variation of experimental parameters to fully explore sources of systematic error (9),
- Obtaining stable empirical regularities,
- Singling out experimental parameters/conditions indispensable for producing such regularities,
- Formulation of experimental arrangements involving only these indispensable parameters so as to present the regularity with clarity.

One way to understand what the process of systematic variation of parameters may look like could be to evaluate computer simulations as a special case of experiments (46). A computer simulation can be defined as a computer program that is used to explore the approximate behavior of a mathematical model. Each instantiation of a simulation depends on a fixed set of initial conditions and parameters. Many simulation experiments involve systematically varying parameters so that the outcomes can be observed for a wide range of conditions. This way, the behavior of the target system may be mapped on the parameter space and the conditions necessary to generate particular patterns of results can be identified. Any specific configuration of a given scientific experiment is akin to a single condition in a computer simulation where parameters are fixed, often at arbitrary values. To reduce dependence on such arbitrariness, scientific experiments in general would benefit from a similar process of exploration to map regularities across the parameter space. Unlike *in-silico* experiments, the process of exploratory experimentation does not or cannot vary experimental parameters all at once (in large part though not exclusively due to resource constraints) nor can it always vary all parameters. In exploratory experimentation, the (necessarily constrained) parameter space is explored iteratively over time but still ultimately aims to identify the conditions necessary to generate particular patterns of results.

[19] Although the standard view may not necessarily preclude the interconnectedness of experiments, it limits the role of experiments to theory testing rather than exploration. In this framework, a network of experiments could very well be designed to test different predictions of a well-specified theory and evaluated in light of that theory, collectively providing a strong test for the theory. Exploration, too, may be theory informed, however, its primary aim is not to test theoretical predictions (45). The interconnectedness of the experiments is an integral part of the process of iterative, exploratory experimentation whereas the standard view assigns each individual experiment a more critical epistemic role.

Epistemic iteration. Exploratory experimentation can also be seen through the lens of *epistemic iteration* as introduced by Hasok Chang in *Inventing Temperature* (47). While investigating the arduous history of understanding and measuring temperature, Chang observes a paradoxical picture of scientific progress where progress can be made by correcting earlier standards, which the new standards were derived of: "What we have is a process in which we throw very imperfect ingredients together and manufacture something just a bit less imperfect." (47, p. 226). Epistemic iteration then is a process of inquiry by which scientific knowledge claims and outputs are repeatedly examined and progressively refined toward achieving certain epistemic goals. Chang's case study documents several productive periods of exploratory experimentation wherein temperature measurement standards were established and revised in an iterative fashion, in the absence of preestablished standards to serve as frames of reference and independently from theoretical progress made elsewhere. Chang observes that "if epistemic iteration works out, we may hope to reach a practical convergence, in which the changes get smaller and subtler as we go on and the system becomes quite stable after a while" (48, p. 233).

Similarly, exploratory experimentation seeks to vary experimental parameters to reach empirical convergence where changes in observed values get smaller and subtler as we navigate the parameter space to locate where stable empirical regularities are produced.

Methodological iteration. In addition to epistemic iteration, exploratory experimentation can also benefit from methodological iteration—a process by which scientists move back and forth between different forms of research practices (49) such as hypothesis testing, instrument development, model selection and refinement. Using the components of our idealized experiment, if epistemic iteration involves iterating between background assumptions and models, methodological iteration would involve iterating between pre-experimental methods and post-experimental statistical methods. Methodological iteration may be necessary in aligning experimental and statistical models closely over time particularly in the absence of well-defined theoretical models (see 33, for a complete description of an iterative process among these three types of models). Methodological iteration typically plays a supporting role in the process of refining knowledge claims as methodological pluralism and triangulation can help eliminate methodological assumptions that are irrelevant to the claim being investigated (12, 50).

Microreplications. Elliott (49) suggests that repetition is central to all forms of iteration. In this sense, different kinds of iterations will repeat different components of an idealized experiment rather than the whole experiment. One particular form of idealized experiment that emerges as a result in exploratory experimentation is microreplications. Guttinger (51) defines microreplications as replication experiments in which an aspect of a previous experiment is repeated as (negative or positive) control condition in a subsequent experiment. In other words, certain components of a previous ex-

periment are exactly copied in the control condition of a subsequent experiment and systematically varied in the treatment condition(s). Microreplications may be used for confirmatory purposes, however, their true potential lies in systematically varying experimental parameters in a stepwise fashion over a sequence of exploratory experiments.

Elementary experiment. One of the possible outcomes of the process of iterative exploratory experimentation, as indicated earlier, is the identification of "pure" or "simple" experiments. Steinle (30, p. S68) provides Faraday's "truly elementary experiment" (52, p. 405) as an example of such a pure case. Before arriving at his truly elementary experiment regarding electromagnetic induction, Faraday performed a series of exploratory experiments, only granting the final experiment the "elementary" status because it exclusively depended only on a set of indispensable experimental parameters. Steinle (30, p. S68) writes: "In the series of experiments preceding the "elementary" one, Faraday systematically varied a lot of parameters of the arrangement such as the direction of motion (relative to the magnetic dip), the mode of motion (e.g., various parts of the circuit or the circuit in its entirety), the form of the circuit, and so on. As a result, he learned which experimental conditions are indispensable in order to bring about an induction effect. The "elementary" status of the experiment depended on those and only those conditions being involved. It shows with particular clarity the general rule of induction of currents by magnets." Steinle further indicates that "in sharp contrast to both Ampère's attraction experiment and Faraday's induction ring experiment, considerations on the nature of the effect or of electricity or of magnetism did not play a role, neither in the design nor in the evaluation of the experiment." In other words, minimal, if at all, theoretical assumptions or preconceptions were needed to design or evaluate the elementary experiment. This appears to be a special form of minimalism where as many theoretical and auxiliary assumptions as possible are removed. Whatever cannot be eliminated defines the most basic form of the experiment that is capable of generating the sought-after regularities.

Schickore (this volume) provides a different example of a process of exploration in viper venom experiments conducted in the 18th century. In particular, the series of experiments published by Felice Fontana stand out. Fontana focused on carefully varying experimental parameters to identify and rule out sources of error. Via diversification and rigorous exploration, Fontana then was able to reduce the number of circumstances the experimental outcomes depended on—similar to Faraday's elementary experiment.

All of this does not mean that exploratory experimentation or its outcomes are completely free from theory (45, 53). Some level of background knowledge is always needed to design, execute, and evaluate experiments as contained in our definition of idealized experiment, even if it simply means using existing experimental paradigms, some known methods, and analytical techniques rather than implying the existence of specific theoretical predictions. As Steinle (30, p. S70) puts it, exploratory experimentation "is driven by the elementary desire to obtain empirical regularities and to find out proper concepts and classifications by means of which those regularities can be formulated" without necessarily being in service of a well-formed theory, although scientists may use specific experimental techniques and instruments developed based on specific theories. Such theories can be used to justify inferences about the target systems made using these techniques and instruments (54). If an elementary experiment is identified as an outcome of exploratory experimentation, however, it will have achieved a minimal degree of dependence on a specific theoretical background or conceptual paradigm (29) by iteratively eliminating the assumptions not necessary for producing the empirical regularity of interest.

This brings us to the final step of our argument: In the absence of well-formed theories, an "elementary" experiment identified via rigorous exploratory experimentation is considered replication-ready.[20]

Minimum viable experiment to replicate

We define the set idealized experiments characterized exclusively by a minimum number of indispensable parameters as the Minimum Viable Experiment (MVE) to replicate[21]. In marketing scholarship and practice, a Minimum Viable Product (MVP) is the version of a new product that is developed with a minimally sufficient, must-have set of features that can be launched quickly to a small group of customers who can be identified as early adopters (55). This minimal version of the product then is reformulated with feedback from initial users and can be realized in many variations of the product. We do not mean this similarity as a full fledged analogy as much as a simple hat-tip to the first author's academic background. The resemblance between MVE and MVP ends at the minimalism. While MVP is part of the *lean startup* process and relies on subsequent testing for designing a final version of the product, MVE is the product of a long period of exploratory experimentation. Nonetheless, we believe the coinage captures the key elements we aim to communicate: minimalism in assumptions and viability for exact replications.

Using our earlier definition of ξ, we next discuss what the minimalism required by MVE would look like. We revisit components of ξ, along with a running example. For generality and simplicity, we use the observations from a coin toss

[20] As we explain later, this statement is not meant to imply that such replications would necessarily be informative. To the contrary, we argue that replications are meaningful only in a limited diagnostic sense and whenever they can be clearly interpreted, the epistemic gains to be made from them are limited. Elementary experiments are such a candidate of a type of experiment whose replications can be clearly interpreted. On the other hand, their replication may not be of particular scientific import.

[21] We opted for the acronym MVE instead of, say, MVER because we believe an MVE is a unique form of idealized experiment that captures the nature of "elementary experiments" in a formalized framework. This minimal version of an idealized experiment would be viable for achieving different scientific aims aside from replication, such as establishing an empirical regularity or identifying a generalized experimental paradigm. Hence the entity we define is not specific for replications; rather, replicability is one of its properties.

experiment with the goal of describing a fair gamble.

- **Background knowledge (K):** Minimum amount of background theory and context that is needed to generate regularities and to be evaluated with ease and clarity. For example, if an experimental paradigm is typically associated with a particular theoretical framework but its value or interpretation does not depend on that framework, these theoretical assumptions can be dropped and the paradigm can be divorced from initially assumed theoretical associations.

  Ancient Romans used coin toss as a game of chance referred to as "heads or ships" where one side of the Roman coin featured a two-headed god Janus on one side and a ship's prow on the other. We can imagine that Roman citizens might have perceived this gamble worked only with their coin, maybe even assumed that the Roman coin had some specific properties (such as shape, material, weight, the face of a god on one side) to make the gamble fair. Imagine a traveling gambler who collects coins from different denominations and performs experiments finds that the material and weight of the coin does not affect its fairness, concluding Roman coins were not necessary for a fair game. As the gambler continues this experiment to test different aspects of the coin's geometry, the MVE will contain only minimal assumptions: the physical symmetry of the coin. Thus the initial assumption that the coin needs to Roman is dropped.

- **Model (M):** Models that require minimal scientific and statistical assumptions. For example, if expected empirical regularities disappear if certain constraints are violated, those constraints need to be obeyed by the model.

  Our Roman gambler may train to toss the coin in a specific way such that it always turns ships. Here, the outcomes of the tosses are dependent beyond the symmetry of the coin because the data generating mechanism has been biased. The MVE for a fair game requires a minimal assumption of independence of coin tosses, given the coin.

- **Methods (S):** Minimal methodological assumptions to collect data under a variety of experimental conditions and to perform inference using the data collected. For example, some scientific results depend on whether frequentist or Bayesian inference is performed. If and only if the results do not depend on such a choice, MVE will drop any assumptions specific to inferential methods.

  For a given coin, the Roman gambler uses the following methodology to assess its fairness: Toss the coin 100 times and record the number of ships. A rival gambler uses a different methodology: Count the number of ships as the tosses are performed and stop when there is sufficient evidence for fairness. Our two gamblers may arrive at different conclusions regarding the fairness of this coin. The MVE cannot drop this assumption as the methodological procedures affect how the conclusion is drawn (in particular, the decision rule).

- **Data (D):** Indispensable features of the data structure that are needed to capture regularities. That is, if certain type of data is needed to capture the empirical regularity, that needs to be specified. If not, the MVE will allow for different kinds of data.

  Today we do not use the label "heads and ships" anymore. In the USA, it is "heads and tails", in Germany it is "Kopf (head) and Zahl (number)", in Italy it's "testa (head) and croce (cross)", and so on. On a computer, we would use 0 and 1. It does not matter how we label the levels of the variable of interest. What is essential is that it has only two levels. For example, we assume that the probability of a coin landing on its edge is zero. That is an indispensable feature of the MVE.

Any MVE identified for replication purposes further requires **openness** of ξ because components of MVE need to be well-specified and transparently reported to satisfy viability for replication[22]. Based on these specifications, we define the MVE as follows.

> **Definition.** *Minimum Viable Experiment to replicate*. Let Ξ be the set of values that ξ can take on the Cartesian product space of components K, M, S, D, and denote the power set by $\mathcal{P}(\Xi)$. MVE about a phenomenon $\mathbf{P}$ is the largest subset of $\mathcal{P}(\Xi)\backslash\emptyset$ on which $R(d, c)$ evaluates to TRUE for the existence of $\mathbf{P}$.

By this definition, MVE is not necessarily a unique ξ. The defining characteristic of MVE is that it is the union of all ξ that contains the essential (in the sense of minimal) experimental conditions such that the result from each ξ is conducive to exact ξ'.

The process of exploratory experimentation visits many experiments that may lead to scientific discoveries. The parameters of these experiments are often fixed arbitrarily. Only by looking at a whole sequence of experiments can we obtain a complete picture of the phenomenon under study and can pinpoint the conditions that are necessary to generate empirical regularities[23]. This big picture allows us to release some of the assumptions characterizing any specific experiment that are not indispensable to produce a given empirical regularity and iteratively eliminate as many auxiliary hypotheses as possible that come attached to particular experimental configurations. As a result, problems arising from

[22] In Buzbas et al. (16), we examine and specify what level of openness and copying is needed to inform a valid replication experiment. Figure 2 shows how conceptual scope increases as the degree of openness in experimental components decreases. Only when all necessary components conditional on an inferential goal are open, can exact (or trivially non-exact) replication experiments be performed.

[23] The same may at times be true for experiments designed for theory testing but is not necessarily so.

the openness of conceptual scope are eliminated as the MVE determines its boundaries.

MVE specifies the necessary conditions to produce empirical regularities. In theory, an exact replication of MVE is meant to provide evidence with regard to the existence of these regularities, providing evidence that cannot be readily attributed to auxiliary hypotheses, since they will have been meticulously eliminated through rigorous exploratory experimentation. The reproducibility rate obtained by replicating MVE is a valid exact estimate of the reproducibility rate of the phenomenon of interest. In practice, this may be too lofty a goal to achieve as we will discuss in the conclusion section.

MVE is distinct from a standard RCT which is also characterized by minimal assumptions and limited prior knowledge (34). RCTs aim at performing causal inference, extrapolation out of trial samples, and generalization across different contexts and their results are often used to inform social, economic, and public health policies. In practice, many RCTs suffer from open conceptual scope and misspecified causal models. The goal of MVE is to show the existence of regularities rather than to generate causal explanations and estimate an average treatment effect. MVE relies on a series of interconnected experiments to identify sources of error and eliminate assumptions necessary to generate the result, rather than randomization alone. Instead of informing policy, MVE informs conceptual representations and classifications to formulate empirical regularities. Indeed, the iterative process aimed at eliminating assumptions has epistemic importance regardless of it producing an MVE and even when it does, the information gained throughout the process may be more valuable than such a final outcome.

Conclusion

> "Only when certain events recur in accordance with rules or regularities, as is the case with repeatable experiments, can our observations be tested — in principle — by anyone. We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them. Only by such repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility, are in principle intersubjectively testable."—*Popper (56, p. 46)*

We showed that in regular scientific practice, repeatable experiments cannot guarantee that we are not dealing with mere isolated coincidences and reproducibility is not a reliable gauge of true regularities. Oftentimes, replication experiments track idiosyncrasies of experimental configurations more closely than any underlying truth (see 16, 21, 57, for rationale and examples). With regard to randomized controlled trials, Deaton and Cartwright (34) argue that "depending on what we want to discover, why we want to discover it, and what we already know, there will often be superior routes of investigation and, for great many questions where RCTs can help, a great deal of other work—empirical, theoretical, and conceptual—needs to be done to make the results of an RCT serviceable." The same can be said for replication experiments. Even exact replications serve a narrow function in scientific process and may be viable in a limited number of situations.

Here we have provided a preliminary sketch of an argument, identifying a specific scientific path that may produce replication-ready experiments, and argued that exact replications may fulfill their aims even in the absence of theoretical maturity or precision, but only when preceded by a process of rigorous exploration. Formalizing the concepts of replication-readiness and MVE helps us explore theoretical implications of experimental design in simulation experiments to understand the role and limitations of replications (28).

Essentially experimentum crucis and MVE both represent some unattainable scientific ideals, led by theoretical or empirical processes, respectively. The experimentum crucis is a theoretical ideal that represents the limits of empirical gain theoretical precision can get us. MVE, on the other hand, is an empirical ideal that represents the limits of conceptual clarity that could be gained thorough experimental exploration. Neither ideal is meant to directly guide research practice so much as to identify the limits of the scientific enterprise. We have argued that the diagnostic aims expected from replication experiments can only be satisfied at these limits, when we design experiments that are replication-ready either because they are maximally and precisely determined by theory or because they are empirically defined by their irreducible components. It is not the experimentum crucis or MVE that represent too high bars for scientists to reach but the diagnostic ideal itself that is an impossible bar for replication experiments to clear.

All things considered, we do not believe that replications or replication-readiness should be the objective of scientific endeavor. The value of knowing to what extent experiments are (not) ready to be replicated is to inform the allocation of scientific resources where epistemic gain can be maximized and to prevent premature conclusions regarding veracity of scientific claims from getting entrenched. While the MVE identifies an empirical ideal for replication-ready experiment in theory-starved fields, it is not presented as a practical solution to the stagnation exemplified in Tao's scenarios presented in the Introduction. The solution, at least one path to the solution (besides pursuing rigorous theoretical investigation), is embracing the exploratory nature of most experiments (e.g., in social and behavioral sciences), and pursuing exploratory experimentation in a systematic, programmatic, rigorous manner not to reach the limit of the process but to continuously improve conceptual clarity and systems-level understanding. By giving up on the diagnostic properties of individual experiments, we can focus on knowledge accumulation within a network of experiments and triangulation of evidence.

Acknowledgements

BD would like to thank Alan Love for organizing the workshop and Jutta Schickore for insightful conversations. BD

and EOB would both like to thank Alan Love and Sam Fletcher for their insightful comments and thought-provoking questions that helped refine the ideas presented here.

Bibliography

1. Terence Tao. What is good mathematics? *Bulletin of the American Mathematical Society*, 44(4):623–634, 2007.
2. Peter Godfrey-Smith. *Theory and reality: An introduction to the philosophy of science.* University of Chicago Press, 2003.
3. Sarahanne Field, Noah van Dongen, and Leo Tiokhin. Reflections on the Unintended Consequences of the Science Reform Movement. *Journal of Trial Error*, 4(1), may 24 2024. https://journal.trialanderror.org/pub/unintended-consequences.
4. M Baker. 1,500 scientists lift. *Nature*, 533:452–454, 2016.
5. Fiona Fidler and John Wilcox. Reproducibility of Scientific Results. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.
6. Markus I. Eronen and Laura F. Bringmann. The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4):779–788, 2021. doi: 10.1177/1745691620970586. PMID: 33513314.
7. Klaus Oberauer and Stephan Lewandowsky. Addressing the theory crisis in psychology. *Psychonomic bulletin & review*, 26:1596–1618, 2019.
8. Mark J Brandt, Hans IJzerman, Ap Dijksterhuis, Frank J Farach, Jason Geller, Roger Giner-Sorolla, James A Grange, Marco Perugini, Jeffrey R Spies, and Anna Van't Veer. The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50:217–224, 2014.
9. Uljana Feest. Why replication is overrated. *Philosophy of Science*, 86(5):895–905, 2019.
10. Rolf A Zwaan, Alexander Etz, Richard E Lucas, and M Brent Donnellan. Making replication mainstream. *Behavioral and Brain Sciences*, 41:e120, 2018.
11. David Peterson and Aaron Panofsky. Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science*, 51(4):583–605, 2021.
12. Brian D Haig. Understanding replication in a way that is true to science. *Review of General Psychology*, 26(2):224–240, 2022.
13. Jutta Schickore. The significance of re-doing experiments: A contribution to historically informed methodology. *Erkenntnis*, 75(3):325–347, 2011.
14. Bert Baumgaertner, Berna Devezer, Erkan O Buzbas, and Luis G Nardin. Openness and reproducibility: Insights from a model-centric approach. *arXiv preprint arXiv:1811.04525*, 2018.
15. Berna Devezer, Danielle J Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3): 200805, 2021. doi: https://doi.org/10.1098/rsos.200805.
16. Erkan O Buzbas, Berna Devezer, and Bert Baumgaertner. The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, 10 (3):221042, 2023.
17. Aurélien Allard and Simine Vazire. Science needs systematic replicability audits. 2021.
18. Brian A Nosek and Timothy M Errington. What is replication? *PLoS biology*, 18(3): e3000691, 2020.
19. Peter M Steiner, Vivian C Wong, and Kylie Anglin. A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, 227(4):280, 2019.
20. James Bogen. Two as good as a hundred': Poorly replicated evidence in some nineteenth-century neuroscientific research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 32(3):491–533, 2001.
21. Berna Devezer, Luis G Nardin, Bert Baumgaertner, and Erkan Ozge Buzbas. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, 14(5):1–23, 2019. doi: https://doi.org/10.1371/journal.pone.0216125.
22. David Trafimow Amrhein, Valentin and Sander Greenland. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(sup1):262–270, 2019. doi: 10.1080/00031305.2018.1543137.
23. Stefan Schmidt. Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of general psychology*, 13(2):90–100, 2009.
24. Michael Inzlicht. The replication crisis is not over, 2020.
25. Tom Chivers. What's next for psychology's embattled field of social priming. *Nature*, 576 (7786):200–203, 2019.
26. Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han LJ van der Maas, and Rogier A Kievit. An agenda for purely confirmatory research. *Perspectives on psychological science*, 7(6):632–638, 2012.
27. Samuel C Fletcher. Replication is for meta-analysis. *Philosophy of Science*, 89(5):960–969, 2022.
28. Erkan O. Buzbas and Berna Devezer. Statistics in service of metascience: Measuring replication distance with reproducibility rate. *Journal of Entropy*, 26(10):842, 2024.
29. Richard M Burian. Exploratory experimentation and the role of histochemical techniques in the work of jean brachet, 1938-1952. *History and Philosophy of the Life Sciences*, 19(1): 27–45, 1997.
30. Friedrich Steinle. Entering new fields: Exploratory uses of experimentation. *Philosophy of science*, 64(S4):S65–S74, 1997.
31. Duygu Uygun-Tunç and Mehmet Necip Tunç. A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. *MetaPsychology*, 20(X), 2022.
32. Daniel J Simons. The value of direct replication. *Perspectives on psychological science*, 9 (1):76–80, 2014.
33. Berna Devezer and Erkan O. Buzbas. Rigorous exploration in a model-centric science via epistemic iteration. *Journal of Applied Research in Memory and Cognition*, 12(4):583–605, 2023.
34. Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.
35. Friedrich Steinle. Experiments in history and philosophy of science. *Perspectives on science*, 10(4):408–432, 2002.
36. Laura R Franklin. Exploratory experiments. *Philosophy of Science*, 72(5):888–899, 2005.
37. Ronald Laymon. Newton's experimentum crucis and the logic of idealization and theory refutation. *Studies in History and Philosophy of Science Part A*, 9(1):51–77, 1978.
38. Isaac Newton. 1672. new theory of light and colors. *Philosophical Transactions*, 6(80): 3075–3087, 1672.
39. Johannes August Lohne. Experimentum crucis. *Notes and Records of the Royal Society of London*, 23(2):169–199, 1968.
40. Anne M Scheel, Leonid Tiokhin, Peder M Isager, and Daniël Lakens. Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4): 744–755, 2021.
41. Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45, 2022.
42. Paul E Meehl. Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. 1992.
43. Gerd Gigerenzer. Mindless statistics. *The journal of socio-economics*, 33(5):587–606, 2004.
44. Ian Hacking. *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge university press, 1983.
45. Uljana Feest and Berna Devezer. Toward a more accurate notion of exploratory research (and why it matters), January 2025.
46. Claus Beisbart. Are computer simulations experiments? and if not, how are they related to each other? *European Journal for Philosophy of Science*, 8(2):171–204, 2018.
47. Hasok Chang. *Inventing temperature: Measurement and scientific progress.* Oxford University Press, 2004.
48. Hasok Chang. Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. *Philosophical issues in psychiatry IV: Psychiatric nosology*, pages 229–245, 2017.
49. Kevin C Elliott. Epistemic and methodological iteration in scientific research. *Studies in History and Philosophy of Science Part A*, 43(2):376–382, 2012.
50. Remco Heesen, Liam Kofi Bright, and Andrew Zucker. Vindicating methodological triangulation. *Synthese*, 196:3067–3081, 2019.
51. Stephan Guttinger. A new account of replication in the experimental life sciences. *Philosophy of Science*, 86(3):453–471, 2019.
52. Michael Faraday. *Faraday's Diary: Being the Various Philosophical Notes of Experimental Investigation Made by Michael Faraday, During the Years 1820-1862 and Bequeathed by Him to the Royal Institution of Great Britain*, volume 7. G. Bell and sons, Limited, 1932.
53. C Kenneth Waters. The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. *History and Philosophy of the Life Sciences*, pages 275–284, 2007.
54. David Colaço. Rethinking the role of theory in exploratory experimentation. *Biology & Philosophy*, 33(5):1–17, 2018.
55. Eric Ries. Minimum viable product: a guide. *Startup lessons learned*, 3:1, 2009.
56. Karl R Popper. The logic of scientific discovery. 1959.
57. Erkan O Buzbas and Berna Devezer. Tension between theory and practice of replication. *Journal of Trial & Error*, 4(1), 2023.